UNIVERSITÀ DEGLI STUDI DI BERGAMO
DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE
E METODI MATEMATICI°

QUADERNI DEL DIPARTIMENTO

Department of Information Technology and Mathematical Methods

Working Paper

Series "*Mathematics and Statistics*"

n. 2/MS – 2012

A statistical framework for model based air quality indicators and population risk evaluation

by

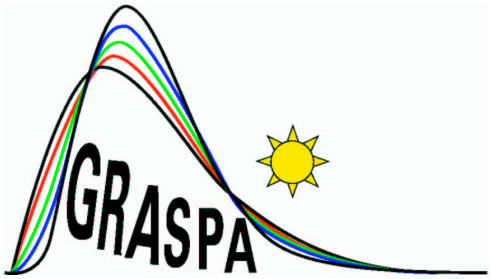
F. Finazzi, M. Scott, A. Fassò

° Viale Marconi. 5, I – 24044 Dalmine (BG), ITALY, Tel. +39-035-2052339; Fax. +39-035-562779

[§] L'accesso alle *Series* è approvato dal Comitato di Redazione. I *Working Papers* della Collana dei Quaderni del Dipartimento di Ingegneria dell'Informazione e Metodi Matematici costituiscono un servizio atto a fornire la tempestiva divulgazione dei risultati dell'attività di ricerca, siano essi in forma provvisoria o definitiva.

ISSN: 2037-7738
GRASPA Working Papers
[online]

Web Working Papers
by
The Italian Group of Environmental Statistics



Gruppo di Ricerca per le Applicazioni della Statistica
ai Problemi Ambientali

www.graspa.org

A statistical framework for model based air quality indicators and population risk evaluation

Francesco Finazzi, Marian Scott, Alessandro Fassò

GRASPA Working paper n.43, February 2012

A statistical framework for model based air quality indicators and population risk evaluation

Francesco Finazzi
University of Bergamo, Italy
E Marian Scott
University of Glasgow, UK
Alessandro Fassò
University of Bergamo, Italy

Summary. This paper is devoted to the development of a statistical framework for air quality assessment at country level and for the evaluation of the ambient population exposure and risk with respect to airborne pollutants. The framework is based on a dynamic coregionalization model which copes with the data complexity and is able to provide high resolution multipollutant dynamic maps. Air quality indicators based on latent variables, exposure indicators and risk indicators are defined at different aggregation levels in space and time and they are evaluated, uncertainty included, on observed air quality data for Scotland for 2009.

Keywords: Air quality indicators; Ambient exposure and risk; Heterogeneous networks

1. Introduction

During the last decades, outdoor air pollution received great attention as one of the main health threats in urban areas, where most of the pollutant sources are located and where pollutant concentration is usually high. Indeed, the air quality problem has been addressed within many research areas, bringing insights into the pollutants' life-cycle and their impact on human health (see Nicolopoulou-Stamati (2005)). Nevertheless, many aspects still deserve to be deepened and new methodologies are needed to answer new complex questions about air quality. The air quality problem is complex in nature due to its typical spatial scale in the order of many kilometers, and to its interaction with the environment and the anthroposphere. Perhaps, the main complication arises from the way airborne pollutants are measured in the field. The high economic costs of installation and maintenance of monitoring networks usually prevent pollutants from being measured with the adequate spatial resolution, resulting in monitoring stations mainly located in critical areas where pollutant concentration and population density are expected to be high. Although this is reasonable from the standpoint of prevention, it may represent a problem when air quality has to be assessed for non-monitored areas. Remote sensing data, characterized by a homogeneous spatial coverage, can mitigate this issue though they must be carefully calibrated with respect to ground-level measurements but may suffer from a high missing data rate (see Fassò and Finazzi (2011)). Alternatively, the approach of Diggle et al. (2010) can be

Address for correspondence: Francesco Finazzi, Department of Information Technology and Mathematical Methods, University of Bergamo, viale Marconi 5, 24044 Dalmine (BG), Italy
E-mail: francesco.finazzi@unibg.it

adopted to jointly model the observed pollutant concentration and the sampling location under preferential sampling.

Another problem related with air quality is how air quality is assessed for large regions, potentially countries. In a way, air quality assessment is the process of deriving a small set of values, reflecting air quality, by analyzing the raw measurements collected by the monitoring networks. When the sets of pollutants measured at different monitoring stations are different, it is not always clear how to define daily and yearly air quality indicators for the whole region or how to evaluate their uncertainty. Moreover, it is not straightforward to compare across years when in each year, the structure of the monitoring network (sites included) and the quantity of missing data differ.

Finally, and more importantly, being able to assess air quality does not mean being able to assess the potential impact of air pollution on population health. Indeed, the ultimate role of air quality assessment should be, on the one hand, to evaluate whether any actions undertaken to improve air quality have been successful or not (see Scott (2007)), and on the other, to provide real-time and long-term population risk and exposure estimation. The measures of risk and exposure are used as a basis for epidemiological studies, where respiratory hospital admissions and possibly other disease rates are correlated with pollutant exposure (see Lee et al. (2009)).

Note that, in this work, the focus will be on the so called ambient exposure rather than on the personal exposure. A review of ambient exposure estimation methods can be found in Jerret et al. (2005) though limited to the intraurban case. A good example of personal exposure estimation, on the other hand, can be found in Zidek et al. (2007). Although not impossible, however, it would be impractical to extend the personal exposure approach from city-size to country-size regions. On the contrary, we aim to provide high resolution ambient exposure maps at the country level, where the exposure is directly represented by the pollutant concentration measured by the monitoring stations. Aware that this may introduce an ecological bias, we point out that the current legislation in terms of pollutant exposure is based on temporal averages of the measured pollutant concentration at the monitoring stations.

The above mentioned aspects related to air quality have usually been considered separately. The main aim of this work is to provide a statistical framework where those aspects can be addressed in a unified way and to provide a set of statistical tools that environmental agencies can adopt in order to handle the current high profile topics of air pollution and health impact, at the country level and with respect to the air quality legislation in force. The framework should be based on a multivariate space-time statistical model, general enough to account for the complex data structure related with the number of pollutants and the way they are measured in space and time. The model itself should be flexible enough to provide results and the respective uncertainty at different levels of aggregation in space and time. The estimation procedures must be stable and efficient both when large datasets are considered and when the datasets are characterized by a high missing data rate and heterotopicity (variables observed at non-collocated sites).

The rest of the paper is organized as follows. Section 2 of this paper is dedicated to the Dynamic Coregionalization Model (DCM) introduced by Fassò and Finazzi (2011). Parameter estimation and space-time pollutant concentration mapping are discussed for multivariate data observed in a heterotopic configuration. In Section 3, the problem of defining aggregated air quality indicators for state-size regions is introduced and a model derived from the DCM is considered. Exposure and risk assessment indicators based on coupling population spatial distribution and model outputs are defined in Section 4. As an

application, air quality data for 2009 collected over Scotland are considered in Section 5 and analyzed by means of the statistical tools developed in this work.

2. The Dynamic Coregionalization Model

Hierarchical models represent a useful statistical approach for the analysis of environmental data and they have been applied profitably both in frequentist and Bayesian contexts. Examples can be found in Banerjee et al. (2004) and Cressie and Wikle (2011) for the univariate and the multivariate case. The Dynamic Coregionalization Model is a hierarchical multivariate space-time model based on latent variables introduced by Fassò and Finazzi (2011). Let $\mathbf{y}(\mathbf{s},t) = (y_1(\mathbf{s},t), ..., y_q(\mathbf{s},t))$ be the $q-$dimensional data response vector at the spatial location $\mathbf{s} \in \mathcal{D} \subset \mathbb{R}^2$ and at time $t \in \mathbb{N}^+$. The general form of the model is the following:

$$\mathbf{y}(\mathbf{s},t) = X(\mathbf{s},t)\beta + K\mathbf{z}(t) + \boldsymbol{\gamma} \odot \mathbf{u}(\mathbf{s},t) + \boldsymbol{\delta} \odot \mathbf{w}(\mathbf{s},t) + \boldsymbol{\varepsilon}(\mathbf{s},t) \tag{1}$$

where $X(\mathbf{s},t)$ is a matrix of known covariates and $\boldsymbol{\beta} = \left(\boldsymbol{\beta}_1', ..., \boldsymbol{\beta}_q'\right)'$ is a vector of global coefficients. The $p-$dimensional latent temporal state $\mathbf{z}(t) = (z_1(t), ..., z_p(t))'$ has the Markovian dynamics

$$\mathbf{z}(t) = G\mathbf{z}(t-1) + \boldsymbol{\eta}(t) \tag{2}$$

with G a stable transition matrix and $\boldsymbol{\eta} \sim N(0, \Sigma_\eta)$. The $q \times p$ matrix K is the loading matrix of known coefficients. The latent spatial component is modeled by both $\mathbf{u}(\mathbf{s},t) = (u_1(\mathbf{s},t), ..., u_q(\mathbf{s},t))$ and $\mathbf{w}(\mathbf{s},t) = (w_1(\mathbf{s},t), ..., w_q(\mathbf{s},t))$ which are i.i.d. over time. For each fixed t, $u_i(\mathbf{s},t)$, $1 \le i \le q$ is a latent zero mean Gaussian process with variance-covariance matrix function $\Gamma_i = cov(u_i(\mathbf{s},t), u_i(\mathbf{s}',t)) = \rho_i(h, \boldsymbol{\theta}_i)$, where ρ_i is a valid correlation function parametrized by $\boldsymbol{\theta}_i$ and $h = \|\mathbf{s} - \mathbf{s}'\|$ is the Euclidean distance between s and s'. On the other hand, $\mathbf{w}(\mathbf{s},t)$ is described by a q-dimensional linear coregionalization model (LCM) of c components

$$\mathbf{w}(\mathbf{s},t) = \sum_{j=1}^{c} \mathbf{w}^j(\mathbf{s},t) \tag{3}$$

where each $\mathbf{w}^j(\mathbf{s},t)$, $1 \le j \le c$ is a latent zero-mean Gaussian process with covariance and cross-covariance matrix function $\Gamma_j^C = cov\left(w_i^j(\mathbf{s},t), w_{i'}^j(\mathbf{s}',t)\right) = V_j \rho_j\left(h, \boldsymbol{\theta}_j^C\right)$, $1 \le i, i' \le q$, $1 \le j \le c$. Each V_j is a correlation matrix and each ρ_j is, again, a valid correlation function. The $\boldsymbol{\gamma} = (\gamma_1, ..., \gamma_q)$ and $\boldsymbol{\delta} = (\delta_1, ..., \delta_q)$ are vectors of scale parameters and $\odot$ is the Hadamard product. Finally, $\boldsymbol{\varepsilon}(\mathbf{s},t) = (\varepsilon_1(\mathbf{s},t), ..., \varepsilon_q(\mathbf{s},t))$ is the measurement error which is assumed white-noise in space and time. In particular, $\varepsilon_i(\mathbf{s},t) \sim N(0, \sigma_{\varepsilon,i}^2)$, $1 \le i \le q$. The parameter set to be estimated is

$$\Psi = \left\{\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\delta}, \boldsymbol{\sigma}_{\boldsymbol{\varepsilon}}^2; G, \Sigma_{\boldsymbol{\eta}}; \boldsymbol{\theta}; \boldsymbol{\theta}^C, \mathbf{V}\right\} \tag{4}$$

where $\boldsymbol{\sigma}_{\boldsymbol{\varepsilon}}^2 = \left(\sigma_{\boldsymbol{\varepsilon},1}^2, ..., \sigma_{\boldsymbol{\varepsilon},q}^2\right)'$, $\boldsymbol{\theta} = \left(\boldsymbol{\theta}_1', ..., \boldsymbol{\theta}_q'\right)'$, $\boldsymbol{\theta}^C = \left((\boldsymbol{\theta}_1^C)', ..., (\boldsymbol{\theta}_c^C)'\right)'$ and $\mathbf{V} = \{V_1, ..., V_c\}$.

It should be noted that the main difference between $\mathbf{u}(\mathbf{s},t)$ and $\mathbf{w}(\mathbf{s},t)$ is that each $u_i(\mathbf{s},t)$, $1 \le i \le q$, is characterized by its own correlation function $\rho_i(h, \boldsymbol{\theta}_i)$ while the LCM imposes, for each of the c components, a unique correlation function across variables. It can be said that $\mathbf{u}(\mathbf{s},t)$ is the direct component while $\mathbf{w}(\mathbf{s},t)$ is the interaction component. Although the model can be estimated with both the components included, preliminary studies suggest that, when real data are considered, it is important to choose one component

or the other depending on the spatial correlation structure of the data. Indeed, the LCM should be included solely when data are known to be spatially cross-correlated while the direct component should be considered solely when the correlation functions describing the q variables are very different. In this latter case, it is still worthwhile to consider the model in its multivariate form since the q variables may be temporally cross-correlated. Finally, if the flexibility of (1) has to be increased, a version of the DCM with spatiotemporal varying coefficients can be considered as detailed in Finazzi and Fassò (2011).

The matrix $Y = Y(\mathcal{S}, \mathcal{T}) = \left(Y(\mathcal{S}_1, \mathcal{T})', ..., Y(\mathcal{S}_q, \mathcal{T})'\right)'$, is the $N \times T$ matrix of all the observations collected at $\mathcal{S} = \{\mathcal{S}_1, ..., \mathcal{S}_q\}$ and time $\mathcal{T} = \{1, ..., T\}$, with $\mathcal{S}_i$ the collection of n_i locations where the variable y_i is observed and $N = \sum_{i=1}^{q} n_i$. The maximum-likelihood (ML) estimate of Ψ is obtained by the expectation-maximization (EM) algorithm as described in Fassò and Finazzi (2011), mostly in closed form with exception of the parameters $\boldsymbol{\theta}$, $\boldsymbol{\theta}^C$ and $\mathbf{V}$. An approximate closed form solution for $\mathbf{V}$ is given in the same work. The whole estimation procedure has been proven to be stable even when large datasets are considered.

Let $\hat{\Psi}$ be the ML estimate of Ψ, then the concentration of the $i-th$ pollutant at a new set of sites $\mathcal{S}_0 \not\subset \mathcal{S}$ and time $t \in \mathcal{T}$ is evaluated by means of a plug-in approach as

$$\hat{\mathbf{y}}_i(\mathcal{S}_0, t) = X_i(\mathcal{S}_0, t)\hat{\boldsymbol{\beta}}_i + K_i \mathbf{z}^T(t) + \hat{\gamma}_i \mathbf{u}_i^T(\mathcal{S}_0, t) + \hat{\delta}_i \mathbf{w}_i^T(\mathcal{S}_0, t) \tag{5}$$

where $\left\{\hat{\boldsymbol{\beta}}_i, \hat{\gamma}_i, \hat{\delta}_i\right\} \in \hat{\Psi}$, $\mathbf{z}^T(t) = E_{\hat{\Psi}}(\mathbf{z}(t) \mid Y)$ is the Kalman smoother output, $\mathbf{u}_i^T(\mathcal{S}_0, t) = E_{\hat{\Psi}}(\mathbf{u}_i(\mathcal{S}_0, t) \mid Y)$ and $\mathbf{w}_i^T(\mathcal{S}_0, t) = E_{\hat{\Psi}}(\mathbf{w}_i(\mathcal{S}_0, t) \mid Y)$ are the estimated latent spatial variables, $X_i(\mathcal{S}_0, t)$ is a matrix of covariates and K_i is again a loading matrix. Note that the Kalman smoother provides a fast algorithm for the evaluation of $E_{\hat{\Psi}}(\mathbf{z}(t) \mid Y)$ using the state-space representation of model (1) (see, for instance Shumway and Stoffer (2006)). On the other hand, the conditional expectations of the latent spatial variables with respect to the observed data are evaluated through the usual formulas of the multivariate normal distribution adapted for the missing data case as detailed in Fassò et al. (2009).

The spatial prediction variance-covariance matrix of $\hat{\mathbf{y}}_i(\mathcal{S}_0, t)$ is given by

$$\Sigma_{\hat{\mathbf{y}}_i}(\mathcal{S}_0, t) = K_i' P^T(t) K_i + A_i(\mathcal{S}_0, t) \tag{6}$$

where $P^T(t)$ is the estimated variance-covariance matrix of $\mathbf{z}^T(t)$ provided by the Kalman smoother and $A_i(\mathcal{S}_0, t)$ is the sub matrix of the co-kriging variance-covariance matrix $A(\mathcal{S}_0 \cup \mathcal{S}, t)$ related to the $i-th$ pollutant and the set of sites $\mathcal{S}_0$.

If the sites in $\mathcal{S}_0$ cover the whole region $\mathcal{D}$ as a fine regular grid, we call $\hat{\mathbf{y}}_i(\mathcal{S}_0, t)$ a map and the ordered collection

$$\hat{\mathbf{Y}}_i(\mathcal{S}_0) = \{\hat{\mathbf{y}}_i(\mathcal{S}_0, 1), ..., \hat{\mathbf{y}}_i(\mathcal{S}_0, T)\} \tag{7}$$

a dynamic map for the $i-th$ pollutant. If, instead of the set of sites $\mathcal{S}_0$, a tessellation $\mathcal{B}$ of the region $\mathcal{D}$ is considered, the change of support problem (see Gotway and Young (2002)) must be addressed and

$$\hat{y}_i(B, t) = E_{\hat{\Psi}}\left(\frac{1}{|B|}\int_{\mathbf{s}\in B} y_i(\mathbf{s})\, ds \mid Y\right) \tag{8}$$

must be evaluated for each block $B \in \mathcal{B}$. However, if the blocks in $\mathcal{B}$ are square pixels whose side length is small compared to the monitoring network site intra-distance, then

$\hat{\mathbf{y}}_i(B,t)$ can be replaced by $\hat{\mathbf{y}}_i(\mathbf{s}^*,t)$, with $\mathbf{s}^*$ the centre of the pixel B. In what follows, the dependence on the estimated parameter set $\hat{\Psi}$ will be dropped in order to simplify the notation.

It must be noted that the dynamic map carries most of the information about the temporal and the spatial dynamics of the ground level pollutant concentration. However, the amount of information is huge and it is rarely useful to decision makers. The following sections describe how aggregate information (uncertainty included) can be derived by considering a different version of the DCM and how dynamic maps can be analyzed in order to evaluate population exposure and risk.

3. Air quality indicators

When environmental space-time data are considered, aggregation can be applied, of course, either over space or time. Although, from a mathematical point of view, both aggregations are equivalent, obtaining aggregated results over time is usually straightforward with respect to spatial aggregation. This is mainly due to the fact that environmental data are collected at regular time steps while the spatial sampling locations are irregularly sparse over the region. Moreover, while the temporal sampling frequency is usually appropriate with respect to the temporal dynamics of the physical phenomenon, the sampling spatial frequency rarely is. For these reasons, the focus of this section is the synthesis of aggregated results over space; in particular, the problem of defining global air quality indicators with measures of uncertainty is addressed.

3.1. Global indicators

With a global air quality indicator, we refer here to a single number able to describe the air quality level for an entire region $\mathcal{D}$ at a specific time t. Since pollutant concentration measurements are never instantaneous, it is implied that the air quality level at time t is the expected air quality level in the interval $(t-1,t]$.

The main role of a global air quality indicator is to provide a way to easily compare disjoint temporal periods with respect to air quality. As already mentioned, air quality reflects the concentration of one or more airborne pollutants. In the ideal case, all the airborne pollutants known to have an impact on human health should be considered. In other words, the problem is usually multivariate.

Before looking for any indicator, it is important to define what the indicator is representative of and which properties it should have. If the region is large, defining a global air quality indicator may be a difficult task since different areas of the region may exhibit different pollutant concentrations due to local conditions.

When air quality is measured at different points in space, a natural choice for the global indicator is the median of pollutant concentration at the sampling sites (see Bruno and Cocchi (2002)). The problem of such an indicator, however, is that it may not be representative of the the region as a whole at time t. Other statistics, such as the mean or the maximum, suffer the same problem. A better approach based on geostatistical modelling can be found in Lee et al. (2011), which also provides measures of uncertainty on the evaluated air quality indicators.

A second problem is the robustness of the indicator with respect to the monitoring sites considered in its evaluation. If the indicator is truly representative of the whole region, then its value should not depend on the particular choice of monitoring stations. This is

not the case if the indicator is defined as simply the mean or the median of the pollutant concentrations. Considering only urban monitoring stations or only rural monitoring stations, for example, gives very different realised indicator values. The problem becomes more prominent when air quality has to be compared across years or across countries and the number of monitoring stations is not constant over time, the monitoring networks are heterogeneous and unbalanced (see Bodnar et al. (2008)) and missing data are present.

In order to solve this impasse, two points are worth mentioning. First of all, different areas of the region are characterized by different concentration magnitudes due to, for example, different urbanization levels. Secondly, assuming a common emission trend over the year for all points of the region, then what really drives pollutant concentration are meteorological conditions. Indeed, when meteorological conditions deteriorate over the whole region (with respect to air quality), then pollutant concentrations increase with the same trend (in a statistical sense) at each monitoring station. This can be easily seen by evaluating the average temporal cross-correlation between sites and noting that it is positive and far from zero.

The above considerations bring us to define a global air quality indicator which is representative of the common variation at monitoring sites rather than of the absolute pollutant concentrations. In other words, if the value of the global air quality indicator doubles from time t to time $t+1$, it means that (on average), the pollutant concentration doubled at each monitoring site from its previous level. The concepts of good and bad air quality are now relative to the average pollution level at the specific site. This reflects the fact that air quality cannot be either bad or good in the same way over the whole region.

Now, from a statistical point of view, the air quality level is considered here as a latent variable which manifests itself through the concentration measurements collected at the sampling sites. Although in a different context, the same idea has been developed by Chiu et al. (2011) in the definition of health factor indices.

In order to estimate the latent air quality level, the following model is proposed:

$$\mathbf{y}(\mathbf{s},t) = K(\mathbf{s})\mathbf{z}(t) + \boldsymbol{\varepsilon}(\mathbf{s},t) \tag{9}$$

which is a reduced and slightly different version of the model described in Section 2. The models do not include any latent spatial variables and the matrix K is now a function of the specific site $\mathbf{s}$. For instance, $K(\mathbf{s})$ can be proportional to the yearly mean pollutant concentration at site $\mathbf{s}$. The dimensionality of $\mathbf{z}(t)$ must be chosen carefully, usually between either 1 or the total number of pollutants q. If the pollutants are known to be highly and positively correlated, then $\mathbf{z}(t)$ can be unidimensional and the global air quality indicator for time t can be easily defined as

$$I_1(t) = z^T(t) \tag{10}$$

where $z^T(t)$ is the estimated latent state output of the Kalman smoother at time t. On the contrary, if the pollutants are not positively correlated or each pollutant has a different effect on population health, then it is better to rely on a q-variate $\mathbf{z}(t)$, in which case each pollutant retain its own temporal trend. In this case, two possible global air quality indicators are

$$I_2(t) = \frac{1}{q}\sum_{i=1}^{q} \pi_i(t) z_i^T(t) \tag{11}$$

$$I_3(t) = \max_{i=1,\ldots,q} \pi_i(t) z_i^T(t) \tag{12}$$

where $z_i^T(t)$ is the $i-th$ component of the estimated latent state $\mathbf{z}^T(t)$ output of the Kalman smoother and $\pi_i(t)$ is a weight reflecting the effect on population health of the $i-th$ pollutant. Note that π_i is time variant, allowing differing weight for the pollutants across time.

When multiple pollutants are considered, the problem of comparing them on a common scale arises. If the above models have to be applied, we strongly suggest to rescale each pollutant with respect to a concentration level L_i, $1 \leq i \leq q$, where L_i may be a specific critical pollutant concentration threshold. This is particularly important when a unidimensional $z(t)$ is used to describe all the pollutants.

With regards to the uncertainty related to the above indicators, this can be evaluated by considering the variance-covariance matrix $P^T(t)$ related to $\mathbf{z}(t)$. In particular, the variance of $I_2(t)$ can be evaluated as

$$Var\left[I_2(t)\right] = \frac{1}{q^2}\sum_{i,j=1}^{q}\pi_i(t)\pi_j(t)p_{ij}(t) \tag{13}$$

where $p_{ij}(t)$ is the $(i,j)-th$ element of the matrix $P^T(t)$. From $\mathbf{z}(t) \mid Y \sim N_q\left(\mathbf{z}^T(t), P^T(t)\right)$, a 95% confidence interval for $I_2(t)$ can be evaluated as $I_2(t) \pm 1.96\sqrt{Var\left[I_2(t)\right]}$. Confidence intervals for $I_3(t)$ do not have, in general, a simple closed form but they can be easily evaluated by considering the quantiles of $N_q\left(\mathbf{z}^T(t), P^T(t)\right)$.

When plotted against time, the indicators I_1, I_2 and I_3 provide an immediate view of the air quality trend over the region considered. This allows comparison across days and years and to test if air quality is either improving or worsening over time. From an epidemiological point of view, however, this kind of information is not rich enough to derive conclusions about the potential impact of pollution on population health, which is the object of the next section.

4. Population exposure and risk assessment

Mapping pollutant concentration over space and time is important to identify critical areas with respect to air quality. In order to evaluate the potential impact of airborne pollution on population health, however, the spatial distribution of the population density must also be considered. Indeed, pollutant concentrations may be high in uninhabited areas, in which case the local impact on population health is either negligible or zero. Of course, most air quality monitoring stations are usually located where critical levels of pollutant concentration may be reached during the year and where the population density is high. Nevertheless, pollutants like ozone are known to be higher in concentration in rural areas, where population density is usually low.

It must be specified that, in this work, pollution is not related to population health in an epidemiological way, namely by correlating pollutant concentrations with clinical data. On the contrary, population exposure and population risk are evaluated by analyzing the interaction between the spatial distributions of the pollutants and the population spatial distribution d for the region $\mathcal{D}$. Exposure and risk indicators are derived in order to compare the potential (or expected) air quality impact on population health across time.

4.1. Exposure indicator

Exposure and risk are related concepts and the respective indicators may carry the same information. However, in particular contexts, exposure and risk might differ substantially. In a way, the risk indicator should reflect the impact of critical (and less frequent) air quality conditions on population health while the exposure indicator should reflects the long-term or the mean effect.

The exposure indicator for the $i-th$ pollutant, the block $B \in \mathcal{B}$ and the temporal frame $\tilde{\mathcal{T}} = \{t_1, ..., t_2\} \subset \mathcal{T}$, $1 \leq t_1 < t_2 \leq T$ is defined here as:

$$\varkappa_i \left(B, \tilde{\mathcal{T}}\right) = \bar{y}_i(B) \cdot d(B) \tag{14}$$

where $\bar{y}_i(B) = \frac{1}{t_2 - t_1 + 1} \sum_{t \in \tilde{\mathcal{T}}} \hat{y}_i(B, t)$, $t \in \tilde{\mathcal{T}}$, is the estimated temporal average concentration of the $i - th$ pollutant while $d(B)$ is the time-invariant population count of block B.

In this case we prefer to evaluate a temporally averaged indicator since, as said before, the exposure indicator should reflects the long term effect. For instance, the set $\tilde{\mathcal{T}}$ can represent a month, a whole season or a year. If needed, the exposure indicator $\varkappa_i \left(B, \tilde{\mathcal{T}}\right)$ can be aggregated over space in order to define the following average exposure indicator for the region $\mathcal{D}$:

$$\varkappa_i \left(\tilde{\mathcal{T}}\right) = \frac{\sum_{B \in \mathcal{B}} \varkappa_i \left(B, \tilde{\mathcal{T}}\right)}{\sum_{B \in \mathcal{B}} d(B)} \tag{15}$$

In order to evaluate how the spatial distributions of population and pollutant concentration interact, an interesting picture is provided by the pollutant concentration density evaluated with respect to population distribution. The pollutant concentration density $h(y)$ over the population can be evaluated by kernel smoothing of the empirical distribution $\hat{H}_i(y)$ of the estimated concentration $\hat{y}_i(B, t)$, namely

$$\hat{H}_i(y) = \frac{1}{|\mathcal{B}|} \sum_{B \in \mathcal{B}: \hat{y}_i(B,t) \leq y} d(B) \tag{16}$$

where $|\mathcal{B}|$ is the number of B in $\mathcal{B}$.

Similarly, for each concentration level L_i related to the $i - th$ pollutant, the time series of the number of people receiving a "dose" higher than L_i can be evaluated as

$$\hat{C}_i(t) = \sum_{B \in \mathcal{B}: \hat{y}_i(B,t) \geq L_i} d(B) \tag{17}$$

4.2. Risk indicator

The risk indicator is defined here by considering a concentration threshold L_i for the $i - th$ pollutant above which the impact on human health is known to be significant. The threshold L_i can be, for example, the concentration level which causes respiratory hospital admissions to increase with respect to a baseline rate. The two risk indicators proposed are:

$$r_i(B, t) = P_{\hat{\Psi}}(y_i(B, t) > L_i) \cdot d(B) \tag{18}$$

$$\tilde{r}_i(B) = P_{\hat{\Psi}}\left(\left|\tilde{\mathcal{T}}(B \mid L_i)\right| > M\right) \cdot d(B) \tag{19}$$

where $P_{\hat{\Psi}}(y_i(B,t) > L_i)$ is the probability that the pollutant concentration y_i exceeds the L_i threshold in block B and time t, while $P_{\hat{\Psi}}\left(\left|\check{\mathcal{T}}\left(B \mid L_i\right)\right| > M\right)$, $\check{\mathcal{T}} \subset \mathcal{T}$, is the probability that the number of days for which L_i is exceeded in block B exceeds M, with $\check{\mathcal{T}}\left(B\right) \subset \mathcal{T}$ the set of days for which the exceedance occurs. The risk indicator defined in (19) reflects the current air quality norm, which usually prescribes a maximum number of days the pollutant concentration can exceed a threshold L. The aggregated risk indicator over space is:

$$r_i\left(t\right) = \sum_{B \in \mathcal{B}} r_i\left(B,t\right) \tag{20}$$

The risk indicator of (19) can be evaluated by noting that

$$\left|\check{\mathcal{T}}\left(B \mid L_i\right)\right| \sim \sum_{t \in \tilde{\mathcal{T}}} Be\left(P_{\hat{\Psi}}(y_i(B,t) > L_i)\right) \tag{21}$$

with $Be(p)$ the Bernoulli distribution with parameter p. Since the sum of independent Bernoulli distributions with varying parameter p has no simple closed formula, (21) must be evaluated numerically. For instance, the distribution of $\left|\check{\mathcal{T}}\left(B \mid L_i\right)\right|$ can be evaluated by Monte Carlo simulation.

As a final remark, it is worth noting that the exposure and risk indicators defined above are conditioned on the specific history of the temporal frame $\mathcal{T}$, in particular to the observed covariates and the estimated latent variables $\mathbf{u}$, $\mathbf{w}$ and $\mathbf{z}$. In other words, both indicators have to be applied only in retrospective analysis and they cannot be considered as characteristics of a particular spatial site $\mathbf{s} \in \mathcal{D}$ independent of time.

4.3. Exceedance probability evaluation

Assuming that the population distribution is available with the proper spatial resolution, a key aspect in assessing risk is the evaluation of the exceedance probability $P_{\hat{\Psi}}(y_i(B,t) > L_i)$. It should be noted that this probability involves the real pollutant concentration $y_i(B,t)$ rather than the kriged concentration $\hat{y}_i(B,t)$. Although $P_{\hat{\Psi}}(\hat{y}_i(B,t) > L_i)$ could be easily evaluated by considering the distribution of $\hat{y}_i(B,t)$, we claim that such a probability is too conservative. In fact, it does not take into account any model miss-specification error. Moreover, we are also interested in evaluating a confidence interval for $P_{\hat{\Psi}}(y_i(B,t) > L_i)$, which is not provided by the dynamic kriging. For these reasons, the following procedure is considered:

(a) Given the model in (1), the observation matrix $\mathbf{Y}$ is used to estimate the model parameter set $\hat{\Psi}$;
(b) The leave-one-site-out cross-validation technique is applied and the cross-validation residuals $e_{\hat{\Psi}}(\mathbf{s},t)$, $\mathbf{s} \in S_i$ related with the $i-th$ variable are considered;
(c) Residuals are studentized with respect to the dynamic kriging variance $\hat{\sigma}^2_{\hat{\Psi}}$, namely:

$$\tilde{e}_{\hat{\Psi}}(\mathbf{s},t) = \frac{e_{\hat{\Psi}}(\mathbf{s},t)}{\hat{\sigma}_{\hat{\Psi}}(\mathbf{s},t)}; \mathbf{s} \in \mathcal{S}_i, t \in \mathcal{T} \tag{22}$$

(d) Considering all the studentized residuals $\tilde{\mathcal{E}} = \left\{\tilde{e}_{\hat{\Psi}}(\mathbf{s},t) : \mathbf{s} \in \mathcal{S}_i, t \in \mathcal{T}\right\}$, their cumulative distribution function $F_{\tilde{\mathcal{E}}}$ is obtained by kernel-smoothing.

(e) For each block B and time t, the exceedance probability is evaluated as

$$P_{\hat{\Psi}}(y_i(B,t) > L_i) \equiv 1 - F_{\tilde{\mathcal{E}}}\left(\frac{L_i - \hat{y}_i(\mathbf{s}^*,t)}{\hat{\sigma}_{\hat{\Psi}}(\mathbf{s}^*,t)}\right) \tag{23}$$

with $\hat{y}_i(\mathbf{s}^*,t)$ the kriged pollutant concentration under the estimated model with parameter set $\hat{\Psi}$. The approximations $\hat{y}_i(B,t) \simeq \hat{y}_i(\mathbf{s}^*,t)$ and $\hat{\sigma}^2_{\hat{\Psi}}(B,t) \simeq \hat{\sigma}^2_{\hat{\Psi}}(\mathbf{s}^*,t)$ are considered negligible since it is assumed that $\mathcal{B} \ni B$ is a fine tessellation of the region $\mathcal{D}$.

Note that the cross-validation residuals are presumed to take into account the model miss-specification error and they are characterized by a higher variance with respect to classical residuals. The transformation at step 3. above is not a real studentization (Cook (1982)) since $\sigma^2_{\hat{\Psi}}(\mathbf{s},t)$ is not an estimate of the residual standard deviation $\sigma^2_e(\mathbf{s},t)$. However, $\hat{\sigma}^2_{\hat{\Psi}}(\mathbf{s},t) \propto \sigma^2_e(\mathbf{s},t)$ and the studentization procedure is applied here in order to homogenize the model residuals which, on their own, are not homoscedastic with respect to space. Indeed, the c.d.f. $F_{\tilde{\mathcal{E}}}$ can be evaluated by considering all the studentized residuals provided they are white noise both in space and time.

4.3.1. Confidence intervals

In order to evaluate if the probabilities given by (23) are reliable, we aim to provide confidence intervals. Let $P_{L_i}(B,t) = P_{\hat{\Psi}}(y_i(B,t) > L_i)$, the γ-level confidence interval for $P_{L_i}(B,t)$ is denoted by $(P_{1,L_i}(B,t), P_{2,L_i}(B,t))$.

The idea here is to consider the asymptotic distribution of the estimated parameter set $\hat{\Psi} \sim N(\hat{\Psi}, \mathfrak{I}^{-1})$, with $\mathfrak{I}$ the Fisher information matrix (see Fassò et al. (2009)), and to sample from it in order to generate a collection of parameter sets $\mathbf{\Psi} = \left(\Psi^{(1)}, ..., \Psi^{(R)}\right)$. For each parameter set $\Psi^{(j)}$, new cross-validation residuals $e_{\Psi^{(j)}}(\mathcal{S}_i,t)$ are evaluated and the procedure of the previous paragraph is applied. This allows a sample of exceedance probabilities

$$P_{\mathbf{\Psi}}(B,t) = \left\{P_{\Psi^{(1)},L_i}(B,t), ..., P_{\Psi^{(R)},L_i}(B,t)\right\} \tag{24}$$

from which to evaluate an approximate confidence interval to be gathered. Note that the assumption $e_{\Psi^{(j)}} \sim e_{\hat{\Psi}}$ can be considered in which case the computation time may be greatly reduced by avoiding, for each $\Psi^{(j)}$, the leave-one-site-out procedure.

5. The Scottish case

The methodology and the statistical tools discussed in the previous sections are applied here to Scottish air quality data for the year 2009. The air pollution standards and the air quality objectives considered in the analysis are based on the Air Quality Standards (Scotland) Regulations 2007 for the purpose of Local Air Quality Management. A summary of the current standards and objectives can be found in DEFRA (2009).

This section is organized as follows. Paragraph 5.1 describes the data considered in terms of pollutants, population distribution and covariates. Paragraph 5.2 reports the model estimation results and some examples of dynamic maps. The global air quality indicator for Scotland is evaluated in paragraph 5.3 while the population exposure and risk indicators are developed in paragraph 5.4.

5.1. Data description

The data sources considered in this work are essentially three: the ground-level concentration of airborne pollutants, measured by the Scottish Automatic Urban Network, the population spatial distribution downloaded from the Oak Ridge National Laboratory and the meteorological covariates downloaded from the NASA Global Modeling and Assimilation Office. Each data source is characterized by a different spatial and temporal resolution, as described hereafter.

5.1.1. Pollutant concentrations

The Scottish Automatic Urban Network provides hourly mean data on six main airborne pollutants, namely nitrogen dioxide (NO_2), ozone (O_3), carbon monoxide (CO), sulphur dioxide (SO_2) and particulate matters (PM_{10} and $PM_{2.5}$). In this work, only the NO_2, O_3, and PM_{10} concentration data are considered since they are measured at sufficient monitoring stations to justify a space-time analysis. Moreover, the hourly mean data are averaged in order to work with daily data.

For the year 2009, the number of monitoring stations is 81. Each station measures only a subset of the three pollutants considered and missing data are possible, due to temporary breakdowns of either the station or the single measuring instrument. Days with less than 75% hourly data (18 hours) are considered as days with missing data. The exact number of monitoring stations for each pollutant is reported in Table 1, showing that the network is unbalanced in the sense of Bodnar et al. (2008). The respective spatial distributions are reported in Figure 1, from which it is clear that the monitoring stations are not evenly distributed over Scotland as they are mainly located in the most populated area of Scotland.

For each pollutant, the average temporal cross-correlation (weighted for missing data) between sites is reported in Table 1 and they suggest the presence of a common temporal trend between sites. This, in turn, suggests that the global air quality indicators defined in Section 3 can be representative of the air quality over Scotland.

5.1.2. Population distribution

The population distribution has a twofold role here. It is considered as a time-invariant covariate and it is used to evaluate the exposure and risk indicators discussed in Section 4. The Oak Ridge National Laboratory manages the LandScan^{TM} ambient population count database, currently updates to the year 2008 (see Bhaduri et al. (2007)). The database provides 24 hours average population count over the entire world with 30" $\times$ 30" resolution (approximately $1km \times 1km$). The population spatial distribution for Scotland is reported in Figure 2a, from which it is clear that most of the population is located in the central belt along the Glasgow-Edinburgh parallel. In what follows, it is assumed that the population spatial distribution is error-free and constant over 2008 and 2009.

Table 1. summary statistics of the pollutant concentration data for 2009. Mean and standard deviation expressed in μgm^{-3}

pollutant	**#sites**	**mean**	**std**	**missing**	**cross**
NO_2	66	32.19	23.35	12.7%	0.65
O_3	10	55.80	18.99	12.1%	0.71
PM_{10}	60	16.60	8.58	16.1%	0.70

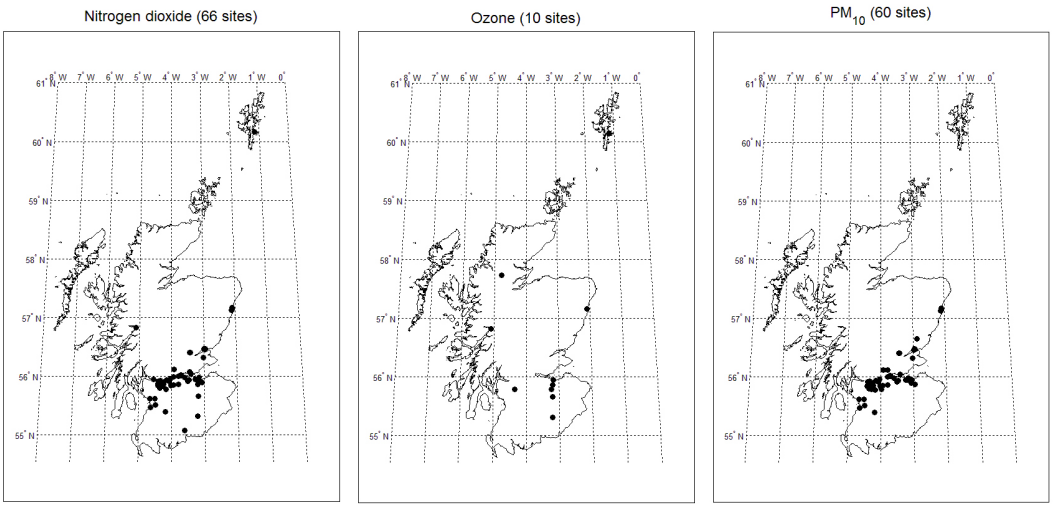


Fig. 1. spatial distributions of the Scottish Automatic Urban Network monitoring sites for NO_2, O_3 and PM_{10}. Number of sites in brackets.

5.1.3. Morphological and meteorological covariates

Pollutant concentrations are known to be related to some anthropological and meteorological covariates due to the physical processes that drive the pollutant diffusion and advection. Covariates can improve the mapping capabilities of the statistical model, reducing estimation uncertainty and providing a better insight into the spatiotemporal pollutant dynamics. The covariates considered in this work are population count (pop), sea level pressure (slp), temperature (t), specific humidity (sh), wind speed (ws) and boundary layer height (blh). Note that the population count is a good proxy of both the pollutant emissions and the site type (urban, sub-urban and rural). The meteorological covariates are downloaded from the NASA Global Modeling and Assimilation Office. In particular, the MERRA (Modern Era Retrospective Analysis for Research and Applications) product (see Rienecker et al. (2010)) is considered, which is characterized by a temporal resolution of one hour and a spatial resolution of $2/3°$ longitude by $1/2°$ latitude. Since the pollutant concentrations are daily averages, the meteorological covariates are also averaged over 24 hours and they are interpolated at 30" $\times$ 30" resolution for mapping purposes.

As a final remark, we point out that we don't claim optimality of the data sources considered in this work. Indeed, the main aim is to provide a statistical methodology for air quality assessment with respect to the legislation in force and to show how it can be applied when routine air quality datasets are considered.

5.2. Model estimation and dynamic mapping

The DCM defined in (1) allows joint modeling of the space-time correlation of all the pollutants considered. However, in order to better define the parametric structure of the multivariate model, it is useful to first estimate as many univariate models as the number of pollutants. In fact, the dimension q of the latent temporal state $\mathbf{z}(t)$, the inclusion of either $\mathbf{u}(\mathbf{s}, t)$ or $\mathbf{w}(\mathbf{s}, t)$ (or both) and the number c of coregionalization components must be

decided before estimating the model. Although models can be compared by means of cross-validation techniques, the number of possible model parametrizations may be large and it is useful to consider the univariate models as a guide to define the parametrization of the multivariate model. As far as the spatial correlation structure concerns, the exponential correlation function has been considered, namely $\rho_i(h, \theta_i) = exp(-h/\theta_i)$, $\theta_i \in \mathbb{R}^+$, $i \in \{NO_2, O_3, PM_{10}\}$

Table 2 reports the value of $\hat{\Psi}$ computed by means of the EM algorithm for each of the univariate models. Note that all the variables and the covariates are log-transformed and standardized. Standardization helps numerical stability and allows direct comparison of the parameter values across pollutants.

By comparing the values of the estimated $\hat{\boldsymbol{\beta}}$ parameters with respect to their standard deviations, it can be seen that all the covariates are significant except $\hat{\beta}_{slp}$ for NO_2. It is worth noting that, though the estimated parameters $\hat{\theta}_i$ may seem very different between each other, their respective standard deviations are quite large. This is particularly true for O_3 which is monitored at only 10 sites. These considerations suggest that a LCM may be appropriate for modeling the spatial latent component, since there is not enough evidence to conclude that the $\hat{\theta}_i$ are different. On the other hand, the $\hat{g}_i$ values are significantly different suggesting that each pollutant should retain its own temporal dynamics, hence $p = 3$ for the multivariate model. The optimal number c of coregionalization components cannot be judged from the results of Table 2 and must be assessed through cross-validation. Cross-validation results not reported here suggest $c = 1$. Table 2 also reports the cross-validation mse (c-mse) obtained by applying the leave-one-site-out technique.

The estimated $\left\{\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\sigma}}^2_{\boldsymbol{\varepsilon}}, \hat{\boldsymbol{\delta}}\right\}$ for the multivariate model are reported in Table 3. The remaining parameters are

$$\hat{G} = \begin{bmatrix} 0.97 & -0.02 & -0.01 \\ -0.17 & 0.87 & 0.11 \\ 0.27 & 0.13 & 0.58 \end{bmatrix}$$

$$\hat{\Sigma}_\eta = \begin{bmatrix} 0.006 & 0.013 & 0.011 \\ & 0.063 & -0.026 \\ & & 0.170 \end{bmatrix}$$

$$\hat{\theta}_1^C = 40.99_{(2.44)}$$

$$\hat{V}_1 = \begin{bmatrix} 1 & -0.79_{(0.06)} & 0.71_{(0.05)} \\ & 1 & -0.61_{(0.09)} \\ & & 1 \end{bmatrix}$$

with the most relevant standard deviations in round brackets. Note that, in this case, the parameter $\hat{\theta}_1^C$ of the spatial correlation function is common for all the pollutants considered. The matrix $\hat{V}_1$ shows that the component of $\mathbf{w}(\mathbf{s}, t)$ related to O_3 is negatively correlated with the components related to the other pollutants. Note also that $\hat{\delta}_i \simeq 0.5$ for each pollutant, namely $\mathbf{w}(\mathbf{s}, t)$ accounts for about half of the data variability. By comparing the c-mse reported in Tables 2 and 3, the gain in terms of prediction capability for the multivariate model can be appreciated. The reduction in the c-mse is particularly evident for O_3 which has the sparsest monitoring network and benefits more from the spatiotemporal correlation with the other pollutants.

The maps of Figure 2 show the yearly average pollutant concentration based on the multivariate model and evaluated through equations (5-6). The yearly averages are obtained

from the estimated dynamic maps $\hat{\mathbf{Y}}_{NO_2}(\mathcal{S}_0)$, $\hat{\mathbf{Y}}_{O_3}(\mathcal{S}_0)$ and $\hat{\mathbf{Y}}_{PM_{10}}(\mathcal{S}_0)$ back-transformed to the original scale, with $\mathcal{S}_0$ given by a regular grid with spatial resolution 30" $\times$ 30" within the Scottish boundaries.

The dynamic maps come with the spatial prediction variance-covariance matrices $\Sigma_{\hat{\mathbf{y}}_i}$ defined in (6). The prediction variance, represented by the diagonal of $\Sigma_{\hat{\mathbf{y}}_i}$, can be seen as the uncertainty on the estimated pollutant concentrations and it provides information about reliable the estimates are. From the point of view of the decision-makers, however, both pieces of information may not be enough to translate the analysis results into actions. For instance, it is not clear how two sites characterized by the same average pollutant concentration but different uncertainty should be treated. The analysis and the examples reported in the next two paragraphs show how to produce useful and immediately interpretable results with respect to the current air quality standards.

5.3. Global air quality indicator

The yearly average maps of pollutant concentration discussed in the previous paragraph can be used to identify (also visually) critical areas with respect to air quality. However, they do not tell the whole story since the temporal dimension is missing. Indeed, those maps can be considered as aggregated data in the sense that each map pixel is a temporal average.

If the aim is to assess the daily air quality over Scotland, extracting useful information by just looking at $T \cdot q = 365 \cdot 3$ daily maps is not feasible. In this paragraph, the global air quality indicators defined in Section 3 are taken into account. Since NO_2, O_3 and PM_{10} are known to have different temporal dynamics over the year, the global air quality indicator defined in (12) is considered. With regard to the estimation of the latent temporal state $\mathbf{z}(t)$, the model defined in (9) is considered with

$$K(\mathcal{S}) = \begin{bmatrix} \frac{\bar{\mathbf{y}}_{NO_2}(\mathcal{S}_{NO_2})}{\bar{y}_{NO_2}} & 0 & 0 \\ 0 & \frac{\bar{\mathbf{y}}_{O_3}(\mathcal{S}_{O_3})}{\bar{y}_{O_3}} & 0 \\ 0 & 0 & \frac{\bar{\mathbf{y}}_{PM_{10}}(\mathcal{S}_{PM_{10}})}{\bar{y}_{PM_{10}}} \end{bmatrix} \tag{25}$$

where $\bar{\mathbf{y}}_i(\mathcal{S}_i) = T^{-1}\sum_{t=1}^{T} \mathbf{y}_i(\mathcal{S}_i, t)/F_i$ is the average scaled pollutant concentration at the sampling sites $\mathcal{S}_i$ for the $i-th$ pollutant, $i \in \{NO_2, O_3, PM_{10}\}$ while $\bar{\mathbf{y}}_i = |\mathcal{S}_i|^{-1}\sum_{\mathcal{S}_i} \bar{\mathbf{y}}_i(\mathcal{S}_i)$. The scaling factors F_i allow the pollutants to be comparable in terms of their impact on population health. The weights in (12) are chosen to be $\pi_i(t) \equiv 1$, $\forall i, t$.

The UK air quality index and banding system (IBS) approved by the UK Committee on Medical Effects of Air Pollution Episodes (COMEAP) is characterized by a 1-10 index divided into four bands, namely low, moderate, high and very high. Each index value corresponds to a concentration range where the pollutant concentration can fall when measured over a period of time ΔT. The limits of each range depend on the particular pollutant as well as ΔT. For PM_{10}, an index value equal to 10 corresponds to a running 24 hour mean concentration ${}_{24h}\bar{y}(\mathbf{s})$ equal or higher to 128 $\mu g m^{-3}$. Although the data considered in this work are daily average concentrations rather than running 24 hour means, it makes sense to consider $F_{PM_{10}} = 128/10 = 12.8$. Note that the division by 10 is introduced in order to keep the indicator I_3 comparable with respect to the index of the IBS. The scaling factor for NO_2 and O_3 are not immediately available since the IBS prescribes ranges for the running 8 hour mean ${}_{8h}\bar{y}(\mathbf{s})$ for O_3 and

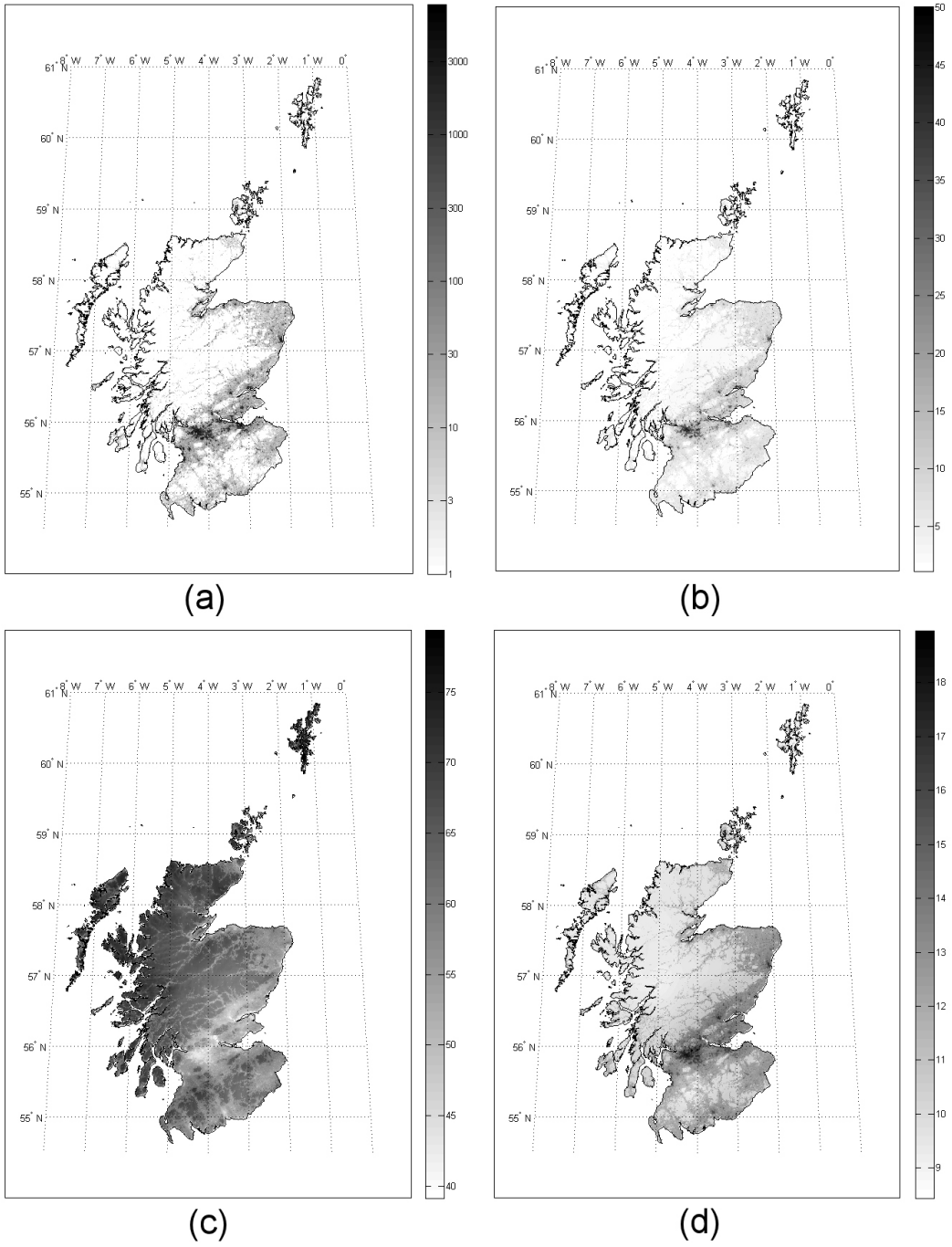


Fig. 2. (a) population spatial distribution - (b) yearly average NO_2 concentration - (c) yearly average O_3 concentration - (d) yearly average PM_{10} concentration. Concentration expressed in μgm^{-3}.

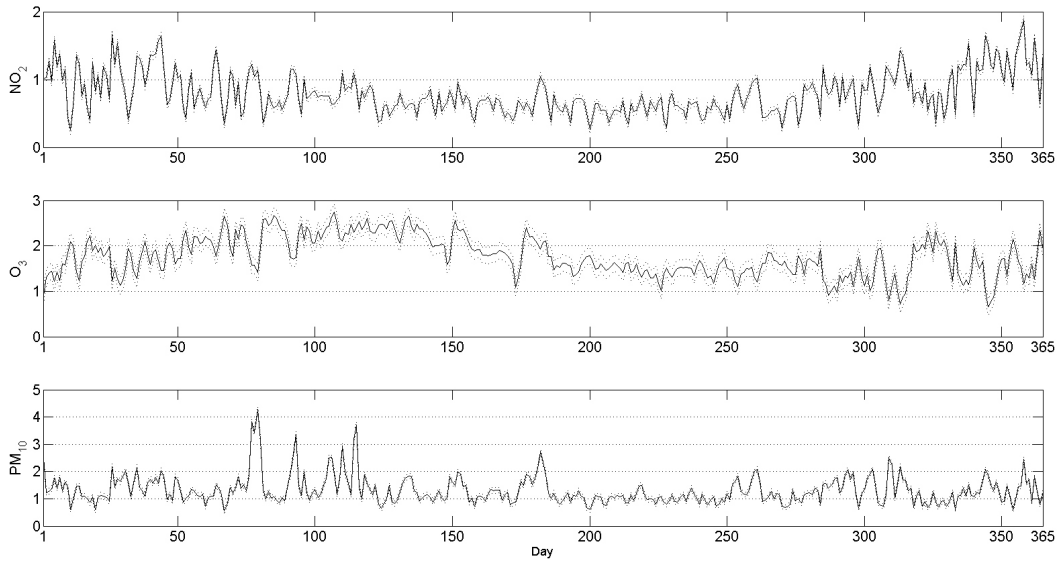


Fig. 3. Kalman smoother output $\mathbf{z}^T(t)$. (top) NO_2 component; (middle) O_3 component; (bottom) PM_{10} component. Error bounds $z_i^T(t) \pm 2\sqrt{p_{ii}(t)}$ as dashed lines.

the hourly mean ${}_{1h}\bar{y}(\mathbf{s})$ for NO_2. Preliminary analysis not reported here suggests that ${}_{24h}\bar{y}_{NO_2}(\mathbf{s}) \simeq \frac{1}{1.91}\max_{day}({}_{1h}\bar{y}_{NO_2}(\mathbf{s}))$ and ${}_{24h}\bar{y}_{O_3}(\mathbf{s}) \simeq \frac{1}{1.15}\max_{day}({}_{8h}\bar{y}_{O_3}(\mathbf{s}))$. The scaling factors are then chosen to be $F_{NO_2} = 764/19.1 = 40.0$ and $F_{O_3} = 360/11.5 = 31.3$, where 764 μgm^{-3} and 360 μgm^{-3} are the concentrations corresponding to an index value equal to 10 in the IBS for NO_2 and O_3 respectively.

After performing ML estimation of model (9), Figure 3 depicts the Kalman smoother output in terms of the estimated temporal component $\mathbf{z}^T(t)$. Error bounds are defined as $z_i^T(t) \pm 2\sqrt{p_{ii}(t)}$. Note that each pollutant is characterized by a different temporal dynamic as expected. Figure 4 shows the evaluated air quality indicator $I_3(t)$ which is representative of Scotland as a whole. By analyzing the temporal series of $I_3(t)$, it can be concluded that, during the year 2009, air pollution over Scotland remained low with the exception of 3 events spread over 7 days during March and April. All the events can be associated with moderate concentration levels of PM_{10} due to adverse meteorological conditions. Note, moreover, that the decisive pollutants are O_3 and PM_{10} while NO_2 is identified in the graph of Figure 4 only three times.

5.4. Population exposure and risk

In the previous paragraph, the global air quality indicator $I_3(t)$ has been considered in order to assess the air quality over Scotland for the year 2009. The use of global air quality indicators like those defined in Section 3 should be encouraged for at least two reasons: they provide an easily interpretable picture of the air quality of a region over time and they can be evaluated rapidly if compared to the dynamic kriging results. On the other hand, if population exposure and risk are to be evaluated, it is important to retain the spatial information on the pollutant concentrations. Indeed, global air quality indicators would be adequate to assess exposure and risk only if the population were distributed uniformly

Table 2. estimated parameters for the univariate DCMs and respective cross-validation mean squared error (c-mse).

	$\hat{\beta}_{pop}$	$\hat{\beta}_{slp}$	$\hat{\beta}_t$	$\hat{\beta}_{sh}$	$\hat{\beta}_{ws}$	$\hat{\beta}_{blh}$
NO_2	0.464	−0.040	0.321	−0.473	−0.197	−0.220
std	0.005	0.021	0.065	0.055	0.015	0.017
O_3	−0.090	−0.229	0.392	−0.297	0.216	0.166
std	0.016	0.026	0.082	0.069	0.018	0.021
PM_{10}	0.121	0.224	0.289	−0.294	−0.084	−0.222
std	0.006	0.038	0.078	0.068	0.020	0.021
	$\hat{\sigma}^2_\varepsilon$	$\hat{g}$	$\hat{\sigma}^2_\eta$	$\hat{\gamma}$	$\hat{\theta}$	c-mse
NO_2	0.367	0.901	0.010	0.463	62.30	0.483
std	0.003	0.038	0.001	0.010	4.28	
O_3	0.274	0.951	0.027	0.427	136.47	0.561
std	0.014	0.019	0.002	0.020	21.96	
PM_{10}	0.286	0.674	0.137	0.516	88.36	0.366
std	0.002	0.054	0.019	0.017	8.91	

Table 3. subset of the estimated parameters for the multivariate DCM.

	$\hat{\beta}_{pop}$	$\hat{\beta}_{slp}$	$\hat{\beta}_t$	$\hat{\beta}_{sh}$	$\hat{\beta}_{ws}$	$\hat{\beta}_{blh}$	$\hat{\sigma}^2_\varepsilon$	$\hat{\delta}$	c-mse
NO_2	0.447		0.309	−0.415	−0.192	−0.211	0.317	0.567	0.439
O_3	−0.166	−0.196	0.368	−0.251	0.208	0.188	0.243	0.451	0.478
PM_{10}	0.089	0.266	0.382	−0.285	−0.064	−0.227	0.244	0.571	0.339

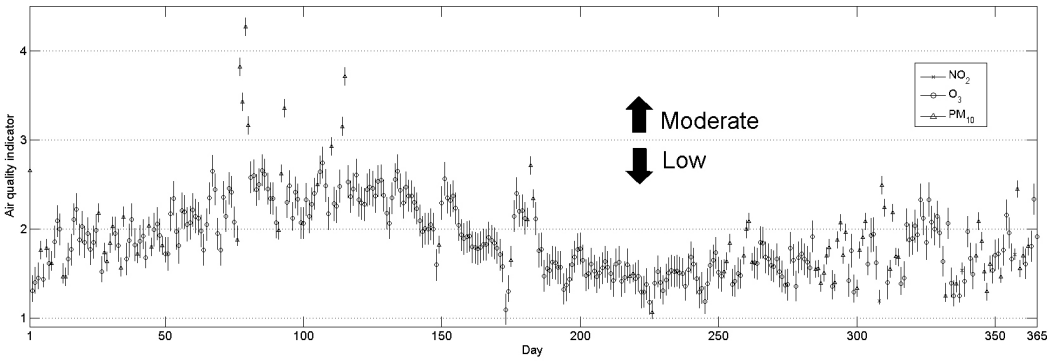


Fig. 4. Scotland daily air quality assessment through the air quality indicator I_3 for the year 2009. 95% confidence interval as vertical lines. Pollutant that give rise to the maximum indicated by the marker.

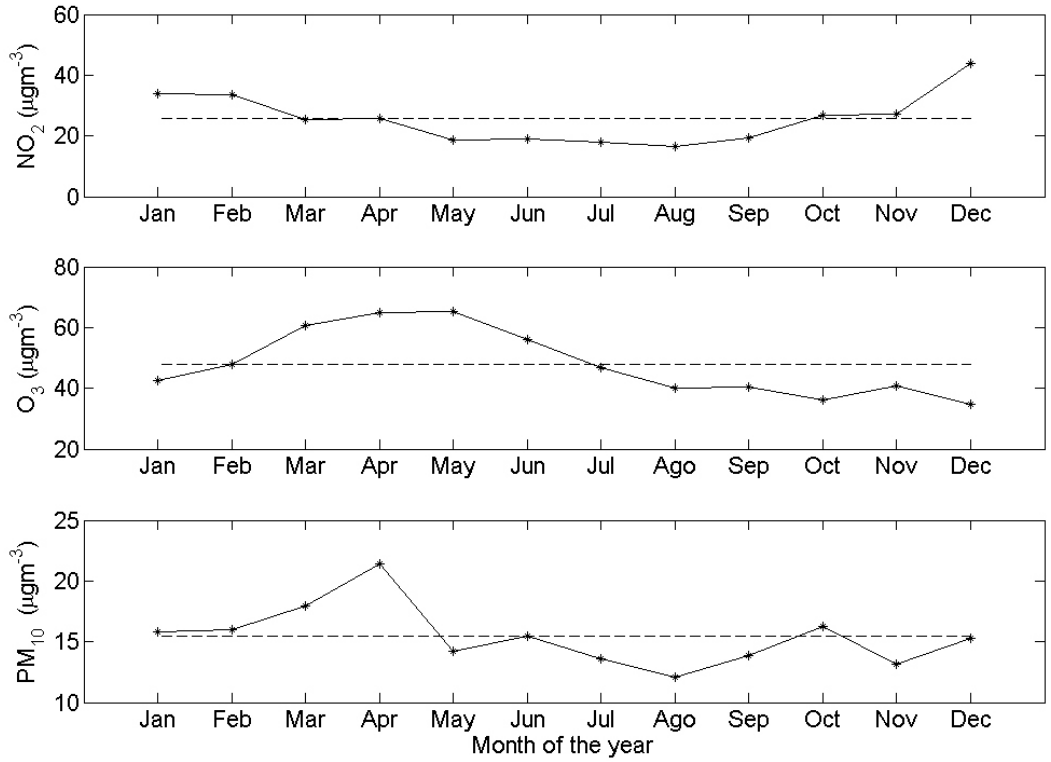


Fig. 5. monthly and yearly (dashed line) average population exposure.

over the region. In this paragraph, population exposure and risk as defined in Section 4 are evaluated by considering the spatial population distribution of Scotland and the collection of dynamic maps estimated for each pollutant.

Figure 5 shows the monthly and the yearly average population exposure evaluated by considering the exposure indicator (15). Note that the exposure values in the graphs can be related to an average Scottish person with respect to where Scottish people live. Figure 6 displays, for each pollutant, the yearly average pollutant concentration density $h(y)$ evaluated by kernel smoothing of (16). By looking at the results of Figure 6, it can be noted that the pollutant concentration density differs greatly across pollutants. In particular, most of the Scottish people share the yearly average PM_{10} concentration while this is not true for NO_2 which is very distributed. Moreover, the pollutant concentration density related to ozone is characterized by a prominent right tail representing people living in rural areas (where the ozone concentration is higher). Note that, although the graphs are reported on the same axis, they are not directly comparable in terms of their health effects.

The daily time series of the number of people exposed to a concentration higher than a threshold (cfr eq. 17) are reported in Figure 7. The thresholds are 105, 87 and 50 μgm^{-3} for NO_2, O_3 and PM_{10} respectively and they have been derived from the Air Quality Standards Scotland Regulations 2007 following arguments similar to those of the previous paragraph. The time series disappear between day 190 and day 315 as a consequence of the fact that the thresholds are never exceeded during July, August, September and October.

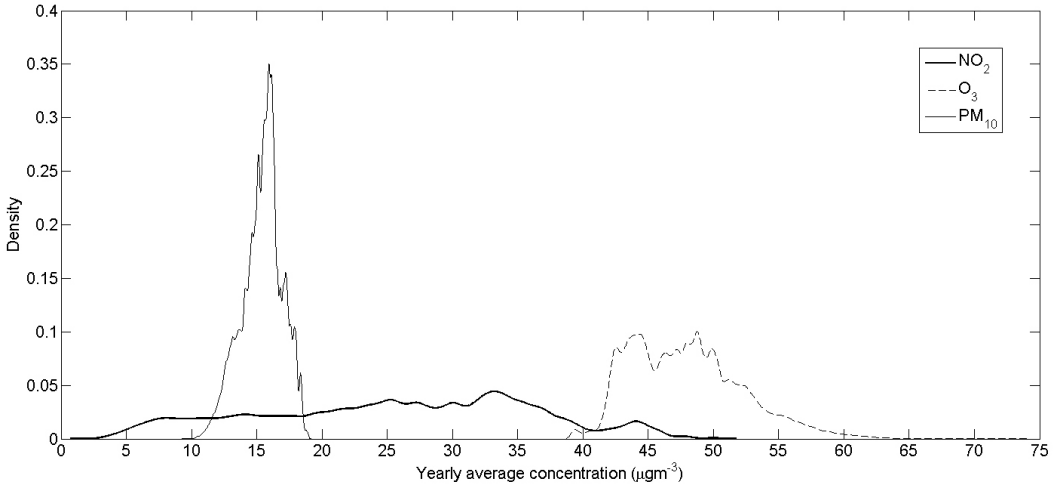


Fig. 6. Kernel estimated yearly average pollutant concentration density distribution with respect to population count.

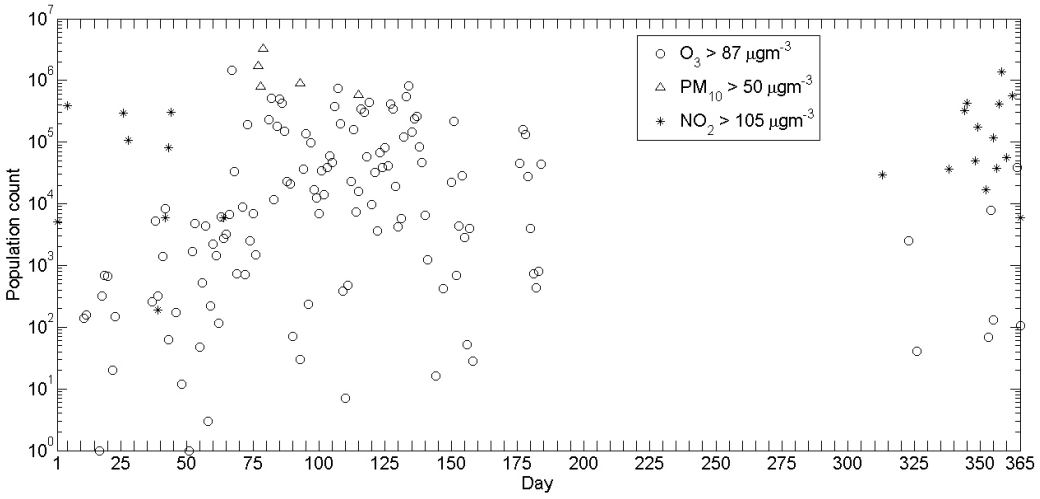


Fig. 7. Daily time series of the population exposed to a pollutant concentration higher than a threshold (thresholds reported in the graph legend).

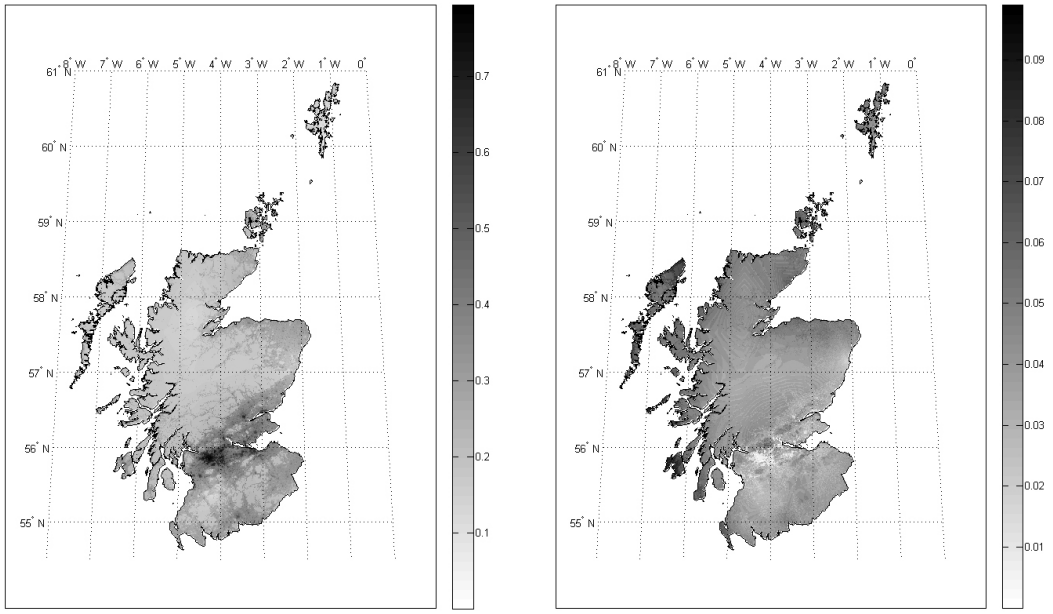


Fig. 8. (left) PM_{10} exceedance probability map with respect to $L_{PM_{10}} = 50\ \mu gm^{-3}$ for March 20, 2009; (right) range of the 95% confidence interval on the exceedance probability.

The exceedance probabilities and their respective confidence intervals have been evaluated for each day and each pollutant at 30" $\times$ 30" of resolution. As an example, Figure 8 reports the exceedance probability map and the confidence interval range map for the 20th of March 2009 (day 79) with respect to the PM_{10} concentration and the threshold level $L_{PM_{10}} = 50\ \mu gm^{-3}$.

The daily exceedance probability maps are used to evaluate the risk indicator defined in (18). As an example, figure 9 shows the daily time series of the risk indicator for ozone and $L_{O_3} = 87\ \mu gm^{-3}$. The 95% confidence intervals has been evaluated following the same arguments of paragraph 4.3.1. With regards to the number of days of exceedance, Figure 10 reports, on the left, the map of the average days of exceedance for PM_{10} (with respect to $L_{PM_{10}} = 50\ \mu gm^{-3}$) evaluated through Monte Carlo simulation, and on the right, the probability map that the threshold $L_{PM_{10}}$ has been exceeded for more than 7 days. Note that the 7 days limit represents one of the objective of the Scotland National Air Quality Strategy to be achieved by 31 December 2010.

Not surprisingly, the probability that the threshold of $50\ \mu gm^{-3}$ has been exceeded for more than 7 days is higher in the Grampian region (north-east) and along the southern border of Scotland rather than in cities such as Glasgow or Edinburgh. However, this is a consequence of the fact that those regions are poorly covered by monitoring stations and the uncertainty on the estimated pollutant concentration is high. The north-west regions are not covered as well but they are characterized by a very low PM_{10} concentration and the exceedance probability is not so high despite the uncertainty.

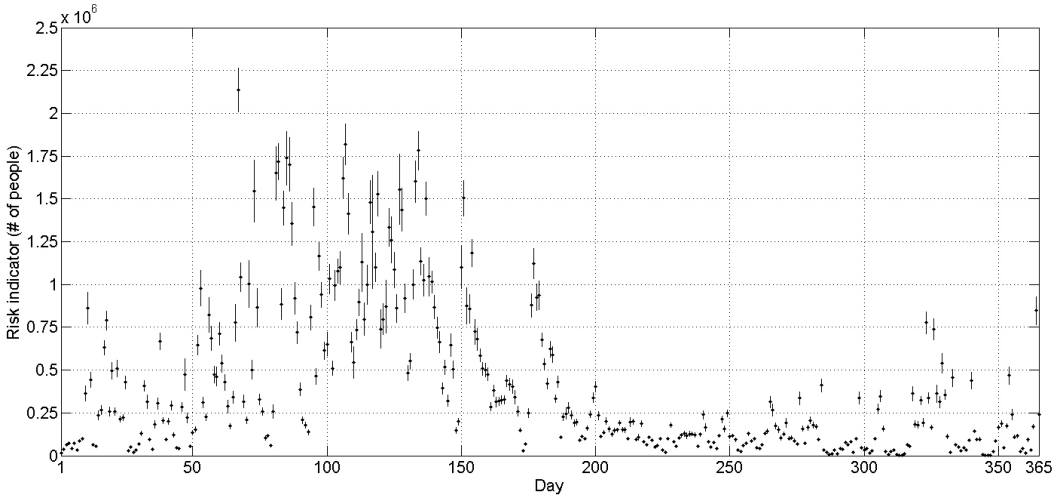


Fig. 9. Ozone daily risk indicator and 95% confidence intervals with respect to $L_{O_3} = 87\ \mu gm^{-3}$.

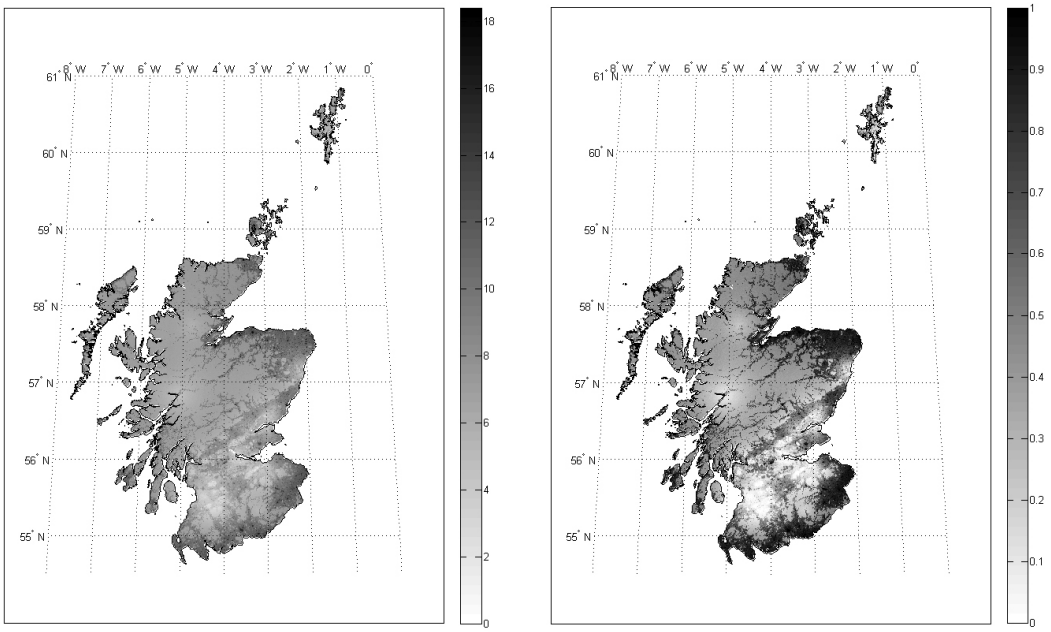


Fig. 10. (left) map of the estimated average days of exceedance for PM_{10} with respect to $L_{PM_{10}} = 50\ \mu gm^{-3}$; (right) map of the probability of exceedance of the 7 days limit.

6. Conclusions

In this paper, the dynamic kriging maps and the respective uncertainty provided by the Dynamic Coregionalization Model have been used as the basis to develop, on the one hand, multipollutant air quality indicators at country level, and on the other, exposure and risk indicators useful to evaluate the impact of pollution on population health.

The DCM is flexible enough to accommodate the data complexity related to ground-level networks characterized by heterogeneous monitoring stations and it naturally copes with the inevitable missing data problem. The indicators are accompanied by measures of uncertainty and they can be provided at different levels of temporal and spatial aggregation in order to study different aspects of the pollution phenomenon.

The DCM and the set of indicators developed represent a complete statistical framework for air quality assessment and management able to assimilate the current air quality legislation and to provide easily interpretable results for decision makers. High resolution exposure and exceedance probability maps provide both an effective way to identify critical areas with respect to air quality and useful information to improve the ground level monitoring network.

References

Banerjee, S., B. Carlin, and A. Gelfand (2004). *Hierarchical Modeling and Analysis for Spatial Data.* Chapman and Hall/CRC.

Bhaduri, B., E. Bright, P. Coleman, and M. Urban (2007). Landscan usa: A high resolution geospatial and temporal modeling approach for population distribution and dynamics. *GeoJournal 69*, 103–117.

Bodnar, O., Cameletti, A. M., Fassò, and W. Schmid (2008). Comparing air quality in italy, germany and poland using bc indexes. *Atmospheric Environment 42*, 8412–8421.

Bruno, F. and D. Cocchi (2002). A unified strategy for building simple air quality indices. *Environmetrics 13*, 243–261.

Chiu, G., P. Guttorp, W. A.H., S. Khan, and J. Liang (2011). Latent health factor index: a statistical modeling approach for ecological health assessment. *Environmetrics 22*, 243–255.

Cook, R. (1982). *Residuals and Influence in Regression.* Chapman and Hall, New York.

Cressie, N. and C. Wikle (2011). *Statistics for Spatio-Temporal Data.* Wiley Series in Probability and Statistics. John Wiley & Sons.

DEFRA (2009). Scottish government local air quality management policy guidance 2009. (available from http://www.scotland.gov.uk/topics/environment/waste-and-pollution/pollution-1/16215/pg09).

Diggle, P., R. Menezes, and T.-l. Su (2010). Geostatistical inference under preferential sampling. *J. R. Statist. Soc. C 59*, 191–232.

Fassò, A. and F. Finazzi (2011). Maximum likelihood estimation of the dynamic coregionalization model with heterotopic data. *Environmetrics*. In printing.

Fassò, A., F. Finazzi, and C. D'Ariano (2009). Integrating satellite and ground level data for air quality monitoring and dynamical mapping. Technical report. GRASPAWorking Paper No.34. www.graspa.org.

Finazzi, F. and A. Fassò (2011). Em estimation of the dynamic coregionalization model with varying coeffcients. Proceedings of Spatial 2, Spatial Data Methods for Environmental and Ecological Processes – 2nd Edition, Foggia, Sept. 1-2, 2011. In printing on Graspa WP, www.graspa.org.

Gotway, C. and L. Young (2002). Combining incompatible spatial data. *Journal of The American Statistical Association 97*, 632–648.

Jerret, M., A. Arain, P. Kanaroglou, B. Beckerman, D. Potoglou, T. Sahsuvaroglu, J. Morrison, and C. Giovis (2005). A review and evaluation of intraurban air pollution exposure models. *Journal of Exposure Analysis and Environmental Epidemiology 15*, 185–204.

Lee, D., C. Ferguson, and R. Mitchell (2009). Air pollution and health in scotland: a multicity study. *Biostatistics 10*, 409–423.

Lee, D., C. Ferguson, and E. Scott (2011). Constructing representative air quality indicators with measures of uncertainty. *J. R. Statist. Soc. A 174*, 109–126.

Nicolopoulou-Stamati, P. (2005). Effects of mobility on health. In *Environmental Health Impacts of Transport and Mobility*, Volume 21 of *Environmental Science and Technology Library*. Springer Netherlands.

Rienecker, M. M., M. Suarez, R. Gelaro, R. Todling, J. Bacmeister, E. Liu, M. Bosilovich, S. Schubert, L. Takacs, G.-K. Kim, S. Bloom, J. Chen, D. Collins, A. Conaty, A. da Silva, and et al. (2010). Merra - nasa's modern-era retrospective analysis for research and applications. *Submitted to Journal of Climate (MERRA Special Collection)*.

Scott, E. M. (2007). Setting and evaluating the effectiveness of environmental policy. *Environmetrics 18*, 333–343.

Shumway, R. and D. Stoffer (2006). *Time series analysis and its applications, with R examples*. Springer, New York.

Zidek, J., G. Shaddick, J. Meloche, C. Chatfield, and R. White (2007). A framework for predicting personal exposures to environmental hazards. *Environmental and Ecological Statistics 14*, 411–431.